The Dimensions of Indexing

W. John Wilbur[†] , MD, PhD and Won Kim[†], PhD

[†] National Center for Biotechnology Information (NCBI)
National Library of Medicine, Bethesda, MD 20894

Indexing of documents is an important strategy intended to make the literature more readily available to the user. Here we describe several dimensions of indexing that are important if indexing is to be optimal. These dimensions are coverage, predictability, and transparency. MeSH® terms and text words are compared in MEDLINE® in regard to these dimensions. Part of our analysis consists in applying AdaBoost with decision trees as the weak learners to estimate how reliably index terms are being assigned and how complex the criteria are by which they are being assigned. Our conclusions are that MeSH terms are more predictable and more transparent than text words.

INTRODUCTION

Keyword indexing is a technique that has been used in many areas to categorize literature for easier and more accurate retrieval. Since texts naturally consist of words, one approach has been to tokenize a text and use some subset of the words obtained in this manner to characterize that text. This is what has become known as automatic indexing and is easily carried out by a simple computer program. A competing approach relies on a controlled vocabulary of keywords which are applied to a text by a human indexer. Each of these methods has advantages and a debate has naturally arisen as to which method should be preferred. In several studies Salton[1, 2] concluded that generally the controlled vocabulary gave a slightly higher precision and the automatic indexing method a slightly higher recall, but the methods were comparable.

Since the work of Salton, others have also compared automatic text word vs. controlled vocabulary indexing. Hersh and Hickman[3] concluded that there was generally no benefit to use of concepts or controlled vocabularies over text words for indexing, but did find a small benefit in the use of MeSH by trained librarians. Yang[4] found an advantage to concepts provided their relations to text words could be automatically learned from a training corpus. Both Srinivasan[5] and Kim, et al.[6] found some benefit when the MeSH was added to text word indexing in a vector retrieval paradigm.

Here we again take up the question of the indexing vocabulary. We propose that a potential indexing vocabulary should be analyzed in terms of three important properties. First, *coverage* denotes the accuracy and ease with which the average query may be expressed within the indexing vocabulary. Clearly the indexing vocabulary forms the basis for a language in which queries must be expressed and documents represented. It must at least be adequate for queries to be expressed in order to serve the purpose of retrieval. Ideally such expressions should be simple to construct. It is clear that text words enjoy an advantage in coverage because they are adequate to express virtually any query. On the other hand controlled vocabulary terms are generally limited in number and may not allow one to express certain queries. However a controlled vocabulary may allow the simpler construction of certain queries because the controlled vocabulary has been constructed to make this possible. For example a query of the form “Heart Failure, Congestive”[MeSH] has over 44 thousand hits in MEDLINE, whereas the phrase “congestive heart failure” occurs in the text of less than 18 thousand MEDLINE documents. While we might approximate the MeSH term with a text query it would be complicated and probably never as accurate at reproducing the concept desired.

The second dimension of indexing which we wish to consider is *predictability*. If one cannot predict when an index term will be assigned, then to that extent one cannot know that it will be useful for retrieval. The level of inter-indexer agreement is a measure of predictability. Funk, et al.[7] report inter-indexer consistencies ranging from 0.3 to 0.6 for different types of MeSH term assignments in MEDLINE. Such a lack of reliability or predictability may seem surprising, but a similar lack of predictability has been observed not only for indexing tasks, but for many other human tasks as well[8-10].

The third dimension of indexing is *transparency*. The easier it is to understand and learn the criteria for applying an index term, the more transparent we consider that term to be. Only those who have learned the MeSH indexing system are able to use it to the greatest advantage[3]. Clearly simplicity and ease of learning, or what we call transparency, is an important ideal for an indexing vocabulary.

In this report we focus on predictability and transparency and compare text words and MeSH terms in MEDLINE on these two scales. To do this we use a state-of-the-art machine learning method to learn how to predict the assignment of an indexing term (either text word or MeSH term) and use a measure of performance at this task as a measure of predictability. We also measure the complexity of the same learning as an estimate of the transparency of the terms. We find that MeSH terms are superior to text words in both predictability and transparency.

METHODS

We chose ten text words and ten MeSH terms for analysis. These words were chosen in all cases to occur in between 3,000 and 3,500 documents in the roughly 12 million document set that comprises MEDLINE. For each such term we performed naïve Bayesian machine learning within all of MEDLINE to learn what documents containing the term are like and applied this to all the documents not containing the term and retrieved the 100,000 documents most closely related to those documents that contained the term. If x represents one of the terms let G_x denote the roughly 3,000 documents that contain the term and let B_x denote the 100,000 retrieved documents that are most closely related but do not contain the term. For each such x we build a database from G_x and B_x. If x is a MeSH term this database contains representations of all the documents in terms of their text words (from titles and abstracts). Stop words are removed, but no stemming is done. Two word phrases without punctuation or stop words are included. No MeSH terms are included. If x is a text word the exact same text word representation of documents is produced with one exception. From the representation of the documents in G_x we remove all occurrences of the text word x. For all x we will denote the resultant database by D_x. Machine learning is applied to D_x to learn how to predict which members of D_x are members of G_x.

Machine learning. AdaBoost is a general machine learning strategy that depends on a machine learning method which is termed the weak learner. AdaBoost iterates the weak learner each time focusing on that part of the training data where the process has not yet succeeded in learning[11]. We use an improved version of Adaboost[12] and boost binary decision trees where the splitting criterion is designed to minimize the error limit computed in Adaboost. We have examined trees of depths 1-5 and have found depth 4 to give the best results and that is what we report here. There is no set limit to the number of rounds of boosting to use in learning. However, in examining many rounds of boosting one generally sees improvement early and then a more or less stable performance. In all cases we repeat the boosting by building another tree till we reach that iteration for which the performance is a maximum within the first 10,000 rounds of boosting. To apply this approach to a D_x we randomly divide D_x by dividing G_x and B_x. Two thirds of G_x and two thirds of B_x are taken for a training set and the remaining one third of each is used as a test set where performance is evaluated.

Evaluation of Transparency. We evaluate transparency in two ways. The simplest evaluation is the number of depth 4 decision trees made, i.e., the number of iterations of the algorithm in reaching peak performance. A somewhat more complicated, but we believe more accurate, method is the following. Let $\{p_i\}_{i=1}^{N}$ be the sequence of precisions obtained over the full set of iterations to the optimum. If for any i $p_{i+1} \le p_i$ drop p_{i+1} from the list. Then for this strictly increasing sequence of precisions define

$$complexity = \sum_{i=1}^{N} \left(p_i - p_{i+1} \right) \log_2 \left(p_{i+1} - p_i \right) \quad (1)$$

This is an entropy and an estimate of how much information is involved in specifying which tree was responsible for correctly categorizing an arbitrary piece of the data that was correctly classified by the final set of decision trees that produce the best observed performance on the test set[13]. While this is somewhat of an average figure because the precisions are averages, it is nevertheless an estimate of how many bits of information would be required to specify the tree involved. If we can assume that all trees are roughly equal in complexity (all are depth 4 decision trees), then the result allows us to compare the complexity of the information captured in learning to predict the different terms. It should be more accurate as an estimate of transparency than simply the number of trees. The lower the complexity, the more transparent the term.

Evaluation of Predictability. Because we use a machine learning method that ranks all the documents in the test set and attempts to rank documents from G_x above documents from B_x, it is convenient to score the result as an 11-point average precision. An 11-point average precision is the average of precisions estimated at the eleven recall values 0%, 10%, 20%, ...,90%, 100%. At each such recall value R the precision is estimated as the highest precision occurring at any rank cutoff where the recall is at least as great as R.

In an attempt to better understand the limitations on predictability we have proposed a model for the assignment of an index term x throughout a document set D assuming D is indexed in equal parts by n different imperfect indexers. For each $k,\ 1 \le k \le n$, let D_k denote that subset of documents from D that would be indexed by k of the n indexers if all indexers indexed all documents in D. Then the sets $\{D_k\}_{k=0}^{n}$ are mutually disjoint and partition the database D. Assuming the documents are assigned randomly to the indexers for actual indexing and only one indexer processes each document, then it is clear that within any D_k only a fraction k/n of the documents receive the indexing term and there is no way to predict which documents are actually assigned the term. Thus in predicting which documents in D_k receive the index term there are two reasonable strategies. Either do not predict the term for D_k or predict the term for all of D_k with precision on D_k of k/n. In general we will not know what D_k is even in the training data because we do not have indexing duplicated n times by n different indexers. If, however, we assume that any indexer on the average assigns the indexing term to a fraction q of the documents and we also assume that indexers do this independent of each other, then we can estimate the size of D_k as a fraction of D. The fraction is just the kth term in the binomial expansion.

$$1=\sum_{k=0}^{n}\binom{n}{k}q^{k}\left(1-q\right)^{n-k}. \tag{2}$$

If we assume as a best case that the D_k are ranked in decreasing order of k, then the precisions for the different sized pieces of D_k can be used to construct a recall-precision curve that is in some sense an upper bound on the predictability possible by any means human or machine. While the independence assumption on which such a computation is based is not strictly true, we believe it is a useful approximation which can be justified on much the same basis that Bayesian retrieval models justify an independence assumption[14, 15]. We have applied this model calculation with different choices of q and n to approximate the recall-precision curves based on the machine learning for index term prediction. We have chosen a representative set to include in our results. The results are not rigorous as upper bounds on predictability, but are only meant to be suggestive approximations of what may be true.

RESULTS

We selected ten MeSH terms in the frequency range described above and ten text words in the same frequency range and chosen to be comparable to the MeSH terms in significance. The two sets of terms are listed in Tables 1 and 2, respectively. An attempt was made to choose the terms as pairs with a reasonably close relationship in order that comparisons between the two sets might be more meaningful. For example the first term in each table is the name of a non-antibiotic drug, the second is the name of a class of antibiotic, and the third a term descriptive of some aspect of the nervous system, etc. The MeSH terms, however, have variations not available with text terms and we have included two starred MeSH terms and two MeSH term-qualifier pairs in our analysis.

We also selected four of the pairs of terms analyzed in the tables and produced recall-precision curves and approximations to these curves coming from equation (2). These curves and their approximations are given in Figures 1 (MeSH) and 2. In all cases we took n to equal 10 and determined a value of q that produced the same 11-point average precision as pertained to the experimental curve. An n of 10 is akin to

Table 1. The data for ten MeSH terms relating to transparency and predictability. The data is based on the machine learning algorithm AdaBoost with decision trees as weak learner.

Mesh Terms	# of iterations	Complexity	11-pap
Atenolol	759	0.435	0.809
Aminoglycosides	14	0.269	0.428
Brain/microbiology	4338	0.577	0.257
Canada/epidemiology	8701	1.782	0.542
Ecology*	3330	0.554	0.299
Fixatives	1881	0.351	0.351
Hemolysis*	13	0.444	0.419
Isomerases	1805	1.572	0.486
Oregon	5639	1.458	0.635
Polyradiculoneuritis	7	0.114	0.519
Average	2649	0.756	0.474

Table 2. The data for ten text words relating to transparency and predictability. The data is based on the machine learning algorithm AdaBoost with decision trees as weak learner.

Text Words	# of iterations	Complexity	11-pap
allopurinol	9394	1.725	0.453
quinolones	1463	0.416	0.356
neuropathology	9996	2.280	0.353
korea	9614	2.180	0.465
environmentally	8694	1.620	0.374
biomarker	9863	1.068	0.212
atelectasis	9831	1.516	0.291
isomeric	9647	1.684	0.312
ohio	8097	2.966	0.628
spasticity	1928	1.326	0.397
Average	7853	1.678	0.384

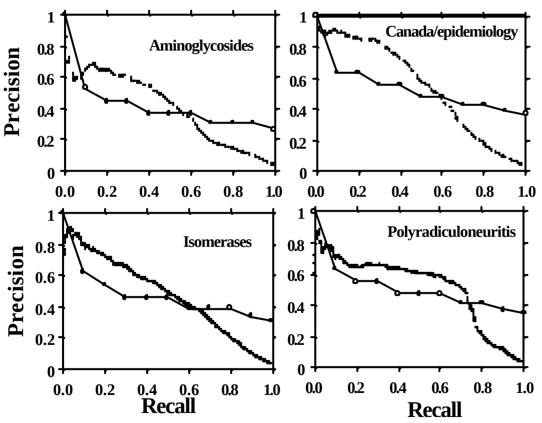

Figure 1. Recall-precision curves for the experimental MeSH data (dark) and approximations based on the independence assumption of equation (2).

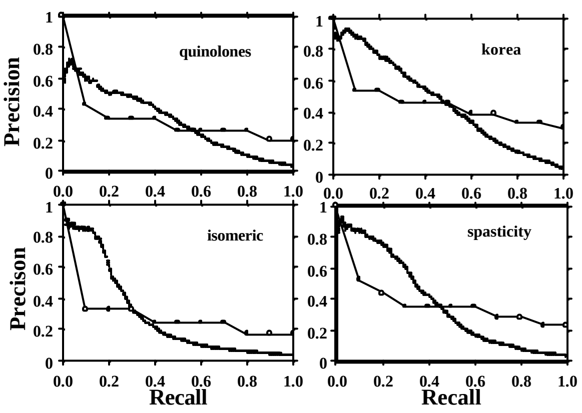

Figure 2. Recall-precision curves for the experimental text word data (dark) and approximations based on the independence assumption of equation (2).

assuming the existence of 10 different criteria by which one could decide whether to assign an index term or not. There is a clear difference between the approximating curves and the experimental curves. The experimental curves are uniformly higher on the left side of the graph and lower on the right side. This suggests that relatively few points tend to fall into the classes D_k with intermediate values of k, but rather to appear in the D_k with k either nearer n or nearer 1 than predicted by (2). This points towards a lack of independence in the data produced by different indexers.

DISCUSSION

It is natural to ask whether it could be valid to base our analysis on MEDLINE entries rather than on the full text documents which they represent. It is true that both the actual MeSH term assignment and the choice of words in the title and abstract constructed by an author depend on the full text. However the title and abstract tell a great deal about an article and it seems likely that valid conclusions can be based on results using just the MEDLINE entry. This is in fact the hypothesis on which previous work has been based[3-6] and we make the same assumption.

Another issue is the choice of the ten pairs of terms. These were not chosen randomly because of the need to pair the terms for similarity of meaning and significance. There could be little validity in comparing a MeSH term and a text word if they did not have similar significance. We selected the pairs by examining the MeSH and text words in the specified frequency range and under the given constraints such pairs are few in number. Bias is of course possible, but it is unclear why it would effect our results.

One may ask if the MeSH and text terms are receiving equal treatment. The text term is left out of the documents in which it occurs to produce the set G_x for learning while no text term is left out of G_x for a Mesh term. One needs to remember that text terms do not correlate with MeSH terms very well generally and that only one text term out of an average of approximately 150 per document is not likely to be significant. Another issue of importance is whether the machine learning method we use provides an accurate approximation to the transparency and predictability of an indexing term. In defense of our method we can point to two facts. First, the AdaBoost algorithm with decision trees as weak learner is one of the best, if not the best, automatic classification method for documents[16-18]. Second, comparison of the results for MeSH and text words shows a strong consistency in the results. With regard to transparency, a comparison of all MeSH term-text word pairs shows that by both number of iterations and by the complexity measure MeSH terms are more transparent than text words. Likewise the same pair wise comparison shows that in 8 out of 10 cases the MeSH terms have a higher 11-pap or predictability. This is true in spite of a positive correlation between complexity and 11-pap within the text word set and no clear relationship between the two concepts within the MeSH term set. This argues strongly that our findings are not noise, but reflect a consistent relationship.

Another aspect of our data that deserves comment is the meaning of the curves in the Figures coming from equation (2). While we do not know what n ought to be, it is clear from the study of Funk, et al. [7] that n is larger than 1. Indexers bring a variety of viewpoints to the indexing task. An n of 10 was chosen as a moderate number for illustrative purposes. The point to illustrate is that if indexers acted independently, then curves of this sort would be upper bounds for the experimental curves, i.e., it would be impossible by any means to predict more accurately than these bounding curves.

The data suggest, however, that indexers are not independent. This basically means that in most cases an index term would either be assigned to a document by most of the indexers or it would be assigned by only a few. A sort of "law of the extremes" applies. The classes D_k still exist and with the same meaning, i.e., the best possible precision on documents in D_k is k/n. But the size of the D_k is not given by equation (2). There still must exist theoretically optimal upper bounding curves but they presumably have a shape more like the experimental curves. We simply need a better theory to determine them.

While our results suggest an advantage for MeSH in transparency and predictability, the situation is somewhat complicated. MeSH has an absolute advantage in predictability because MeSH terms are simply assigned more consistently to documents and one can have more assurance of what will be retrieved by their use as opposed to text words. The greater transparency of MeSH terms also suggests an advantage for MeSH over text words in that the meaning of MeSH terms should be easier to learn. However, text words are more familiar to users so that there is usually less need to learn what they mean and how they are likely to be used in documents.

Finally, there is the issue of coverage. Because there are so many more text words than MeSH terms it is expected that text words would enjoy an advantage in coverage. However, MeSH terms are specifically crafted to represent areas of scientific interest. Thus one cannot simply assume that text words are superior as a language in which to express queries. We hope to investigate the issue of coverage for MeSH as opposed to text terms in future work.

REFERENCES

1. Salton G. Developments in automatic text retrieval. Science 1991;253:974-980.

2. Salton G. A comparison between manual and automatic indexing methods. American Documentation 1969(January):61-71.

3. Hersh WR, Hickam D. Information retrieval in medicine: The Sapphire experience. Journal of the American Society for Information Science 1995;46(10):743-747.

4. Yang Y, Chute CG. Words or concepts: the features of indexing units and their optimal use in information retrieval. In: Safran C, ed. Seventeenth Annual Symposium on Computer Applications in Medical Care. Washington, D. C.: McGraw-Hill, 1994:685-689.

5. Srinivasan P. Optimal document indexing vocabulary for MEDLINE. Information Processing & Management 1996;32(5):503-514.

6. Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proc. AMIA Symp. Washington, D.C., 2001:319-324.

7. Funk ME, Reid CA, McGoogan LS. Indexing consistency in MEDLINE. Bulletin of the Medical Librarians Association 1983;71(2):176-183.

8. Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in human-system communication. Communications of the ACM 1987;30(11):964-971.

9. Saracevic T. Individual differences in organizing, searching, and retrieving information. In: Griffiths J-M, ed. Proceedings of the 54th Annual ASIS Meeting. Washington, D.C.: Learned Information, Inc., 1991:82-86.

10. Swanson DR. Historical note: Information retrieval and the future of an illusion. Journal of the American Society for Information Science 1988;39(2):92-98.

11. Schapire RE. The Boosting approach to machine learning: An overview. MSRI Workshop on Nonlinear Estimation and Classification, 2002.

12. Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Machine Learning 1999;37(3):297-336.

13. Gallager RG. Information Theory and Reliable Commication. New York: John Wiley and Sons, Inc., 1968.

14. Robertson SE, Sparck Jones K. Relevance weighting of search terms. Journal of the American Society for Information Science 1976;May-June:129-146.

15. Yu CT, Salton G. Precision weighting - An effective automatic indexing method. Journal of the Association of Computing Machinery 1976;23(1):76-88.

16. Apte C, Damerau F, Weiss S. Text mining with decision rules and decision trees. The Conference on Automated Learning and Discovery. CMU, 1998:148-155.

17. Carreras X, Marquez L. Boosting trees for anti-spam email filtering. RANLP2001. Tzigov Chark, Bulgaria, 2001.

18. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. The Annals of Statistics 2000;38(2):337-374.